Journal of Multivariate Analysis 107 (2012) 162–168

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

On the border of extreme and mild spiked models in the HDLSS framework

Myung Hee Lee

Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, United States

ARTICLE INFO

Article history:
Received 4 May 2011
Available online 20 January 2012

AMS 2000 subject classifications:
Primary 62H25
34L20

Keywords:
HDLSS asymptotics
HDLSS geometric representation
Principal Component Analysis
Spiked covariance model
Strongly Inconsistent
Subspace Consistent

ABSTRACT

In the spiked covariance model for High Dimension Low Sample Size (HDLSS) asymptotics where the dimension tends to infinity while the sample size is fixed, a few largest eigenvalues are assumed to grow as the dimension increases. The rate of growth is crucial as the asymptotic behavior of the sample Principal Component (PC) directions changes dramatically, from consistency to strong inconsistency at the boundary of the extreme and mild spiked covariance models. Yet, the behavior at the boundary spiked model is unexplored. We study the HDLSS asymptotic behavior of the eigenvalues and the eigenvectors of the sample covariance matrix at the boundary spiked model and observe that they show intermediate behavior between the extreme and mild spiked models.

1. Introduction

Data with more variables than the sample size are emerging in a number of fields, such as microarray experiments, text recognition, and signal processing. Principal Component Analysis (PCA) is widely used for dimension reduction and also for visualization purposes when exploring high-dimensional data. When the dimension is fixed, classical asymptotic results show that the sample covariance matrix provides a good approximation of the population covariance matrix as the sample size n goes to infinity [3,12]. This is no longer the case when the dimension grows comparable to, or even much faster than, the sample size.

Along with the development of new methodologies for high dimensional data, a new family of asymptotics has come with growing dimensionality. In an increasing d scenario, there are different types of asymptotics that have been considered in the literature. A common feature of large dimensional asymptotics is to let the dimensionality and the sample size grow at the same rate. There is a rich body of work on the asymptotics of the sample eigenvalues when the ratio of the sample size to the dimension converges to a positive constant; Bai and Yin [5], Johnstone [11], Bai and Silverstein [4] and Paul [15], to name a few. A study on the sample eigenvalues from a non-identity covariance matrix, particularly, from a spiked population model where a few largest eigenvalues are bigger than the rest is rather recent [11,6,15]. Fan et al. [8] considered estimation of the covariance matrix under some spike structure assumptions on a factor model. It has been known that the spiked model undergoes the phase transition phenomenon where the asymptotic behavior of the eigenvalues and the eigenvectors of the sample covariance matrix changes critically at some boundary of the population covariance model [11,15].

Some researchers have addressed the case where the dimensionality and the sample size grow at a different rate. For example, Portnoy [16] let $n \to \infty$, with d also growing as $n^{1/2}$ in the regression context, Bickel and Levina [7] showed that

E-mail address: mhlee@stat.colostate.edu.

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2012.01.003

Fisher's linear discriminant performs poorly when $d/n \to \infty$, and Fan and Lv [9] reviewed some recent advances in the ultra high dimensional problems where d grows at a non-polynomial rate of the sample size.

The other extreme type of asymptotics, called HDLSS asymptotics, has emerged rather recently [10,14,17]. In [10], they let $d \to \infty$ while keeping the sample size n fixed and the geometric structure of HDLSS data was explored in the time series context where the variables are ρ-mixing. They observed that the data vectors approximately form a regular n-simplex in a large dimensional space. Later Ahn et al. [2] and Jung and Marron [13] extended the results in a multivariate setting where the variables are ρ-mixing under some permutation. In their work, the assumptions on the geometric representation are formulated based on the measure of sphericity,

$$\epsilon = \frac{\left(\sum_{i=1}^{d} \lambda_i\right)^2}{d \sum_{i=1}^{d} \lambda_i^2}.$$

The inequality $1/d \leq \epsilon \leq 1$ always holds and the perfect sphericity of the distribution occurs if and only if $\epsilon = 1$. The geometric representation is achieved if ϵ is not far from the sphericity in the sense that

$$(d\epsilon)^{-1} \to 0 \quad \text{as } d \to \infty. \tag{1}$$

This condition holds for a quite broad range of covariance settings. See [13] for details. It is noteworthy that the assumption (1) is also closely connected to the conditions on the consistency of the PC direction vectors in HDLSS asymptotics. For an intuitive example, consider a family of Gaussian spiked models with varying sphericity where the covariance is a $d \times d$ diagonal matrix with entries $(d^\alpha, 1, \ldots, 1), \alpha \geq 0$. If $\alpha = 0$, the covariance is identity and the larger the parameter α, the larger the perturbation from the spherical Gaussian distribution. In this model, as the dimension increases, the first eigenvalue grows at the rate of d^α whereas the rest of the eigenvalues stay the same. Thus, the parameter α represents the strength of largest eigenvalue signal.

Under the extreme spiked model where $\alpha > 1$, the assumption (1) does not hold. The largest eigenvalue grows so much faster than the rest as d goes to infinity that the covariance tends too be far from the spherical covariance. In other words, the signal for the largest eigenvalue is large enough so that the first sample eigenvector converges to its population counterpart as $d \to \infty$. On the other hand, under the mild spiked model where $\alpha < 1$, the assumption (1) holds and the geometrical representation of the data holds. Simply put, the population covariance is not too far from the identity matrix in the large d limit and the structure of the data points is similar to that of data points generated from the identity covariance. This in turn means that the largest eigenvalue is indistinguishable from the rest in the limit, thus, the inconsistency of the first PC direction is expected. In fact, the first sample PC direction is strongly inconsistent to its population version in the sense that the angle between the two direction vectors converges to $\frac{\pi}{2}$.

As noted by Ahn et al. [2] and Jung and Marron [13], the phase transition phenomenon is also observed in the spiked population model in the context of the HDLSS asymptotics, and the gap between the two regions of the consistency and the strong inconsistency of the PC direction is very thin; $\alpha = 1$. However, the asymptotic behavior of the PC direction at the boundary has not been explored. This work is to fill the gap in the knowledge of HDLSS asymptotic behavior of PC directions in the spiked population model.

In this paper, we consider a broader boundary spiked population model than the Gaussian boundary spiked model above. We address the questions (1) whether some relaxation of geometric representation is possible and (2) whether the signal for the largest eigenvalue is strong enough so that consistency of the first PC direction vector is ensured at the boundary spiked model. We first investigate the HDLSS limit of the sample covariance matrix in Section 2 and study the geometric structure of the HDLSS data vectors in Section 3. The limit behavior of the sample eigenvalues and the sample PC directions are presented in Section 4.

2. The sample covariance matrix

In this section, we investigate the asymptotic behavior of the sample covariance matrices as $d \to \infty$ with the fixed sample size, n. Let $\Sigma_d, d = n+1, n+2, \ldots,$ be a sequence of $d \times d$ covariance matrices. For the sake of simplicity, we drop the notation d from the subscript if the dependence on d is self-evident. Suppose we have a $d \times n$ random matrix $X = [X_1, \ldots, X_n]$ with $d > n$, where $X_j = (X_{1j}, \ldots, X_{dj})^{\mathrm{T}}$ are independent and identically distributed from a d-dimensional multivariate distribution with mean zero and the covariance matrix Σ. Let the eigenvalue decomposition of the covariance matrix be $\Sigma = V \Lambda V^{\mathrm{T}}$, where $\Lambda = \operatorname{diag}(\lambda_{1,d}, \ldots, \lambda_{d,d})$ is a diagonal matrix of eigenvalues $\lambda_{1,d} > \cdots > \lambda_{d,d} > 0$. Consider the sphered data matrix by pre-multiplying the inverse of the square root of Σ, i.e., $Z = \Lambda^{-1/2} V X$. Then Z is a $d \times n$ data matrix and the columns of Z are independent samples from a d-dimensional multivariate distribution with mean 0 and the covariance matrix I. Two different sample covariance matrices are defined without subtraction of the sample means because the population mean is assumed to be zero; the $d \times d$ sample covariance as $S = \frac{1}{n} X X^{\mathrm{T}}$ and the $n \times n$ dual sample covariance matrix as $S_D = \frac{1}{n} X^{\mathrm{T}} X$. Note that these two covariance matrices share the same non-zero eigenvalues.

In what follows, we assume the following:

(A1) The sphered data Z have uniformly bounded fourth moments and are ρ-mixing under some permutation of the variables.
(A2) $\lambda_{i,d}/d \to c_i$ as $d \to \infty$ for some $c_i > 0, i = 1, \ldots, k$ for a fixed $k < n$.
(A3) $\sum_{j=k+1}^{d} \lambda_{j,d}/d \to c_{k+1}$ as $d \to \infty$.
(A4) The ϵ_{k+1}-condition holds, i.e., $(d\epsilon_{k+1})^{-1} \to 0$ as $d \to \infty$, where ϵ_{k+1} is the sphericity measure except for the k largest eigenvalues defined as

$$\epsilon_{k+1} = \frac{\left(\sum_{i=k+1}^{d} \lambda_{i,d}\right)^2}{d \sum_{i=k+1}^{d} \lambda_{i,d}^2}.$$

The assumption (A1) ensures the applicability of the law of large numbers with an increasing d. The assumption (A4) essentially means that, except for the largest k, there are no dominant eigenvalues in the limit. The assumptions (A2)–(A3) capture the essence of the boundary spiked model in HDLSS context where the individual sizes for the largest few eigenvalues grow comparably to the size of the aggregation of the rest of the eigenvalues. This boundary spiked model includes the population covariance matrix, $\Sigma = \mathrm{diag}\,(d, 1, \ldots, 1)$, illustrated in Section 1, but has more flexibility. The signal for the largest few eigenvalues stands out as the dimension grows; however, it does not dominate the sum of the small eigenvalues. They are balanced out in the end, and thus show some intermediate behavior in the HDLSS asymptotics between the extreme and the mild spiked models.

Express the dual sample covariance in terms of the sphered data matrix,

$$nS_D = (V^{\mathrm{T}} \Lambda^{1/2} Z)^{\mathrm{T}} (V^{\mathrm{T}} \Lambda^{1/2} Z) = \sum_{i=1}^{d} \lambda_{i,d} z_i^{\mathrm{T}} z_i,$$

where $z_i, i = 1, \ldots, d$, are the row vectors of Z. Write Z and Λ as block matrices such that

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \qquad \Lambda = \begin{pmatrix} \Lambda_1 & O \\ O & \Lambda_2 \end{pmatrix}, \tag{2}$$

where Z_1 and Z_2 are $k \times n$ and $(d-k) \times n$ partitioned sphered data matrices, and $\Lambda_1 = \mathrm{diag}\,(\lambda_{1,d}, \ldots, \lambda_{k,d})$ and $\Lambda_2 = \mathrm{diag}\,(\lambda_{k+1,d}, \ldots, \lambda_{d,d})$ are $k \times k$ and $(d-k) \times (d-k)$ block diagonal matrices, respectively. The matrices O are the zero matrices with respective sizes. In the following theorem, the limit distribution of the dual sample covariance is achieved.

Theorem 1. *For a fixed n, let $\Sigma_d, d = n+1, n+2, \ldots$, be a sequence of covariance matrices. Let X be a $d \times n$ data matrix from a d-variate distribution with mean 0_d and the covariance Σ_d. Suppose that the assumptions* (A1)–(A4) *hold. Then, the scaled dual sample covariance matrix converges in distribution in the following sense:*

$$\frac{nS_D}{d} \Rightarrow Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n \quad \text{as } d \to \infty,$$

where $C_1 = \mathrm{diag}\,(c_1, \ldots, c_k)$.

Note that the scaled dual sample covariance matrix from the mild spiked model converges to I_n in probability [2,13]. Under the boundary spiked model, we no longer have such a deterministic limit. But the deviation from the identity covariance model is not too large so that the departure from the I_n has a limit distribution. Note that under the additional Gaussian assumption on the data matrix, the dual sample covariance matrix is approximately a shifted Wishart distribution.

3. Geometric representation

In this section, we use the limit behavior of the sample covariance matrix presented in Theorem 1 and study the geometric representation of the data vectors in the large d limit. Under the mild spiked model, the pairwise distances between the data vectors and the norm of the data vectors become approximately deterministic. Moreover, in the large d limit the angle between any two data vectors is approximately $\frac{\pi}{2}$. As a result, the n data vectors approximately form a regular n-simplex in a large d-dimensional space.

At the boundary spiked model, such deterministic geometrical structure is no longer expected since the signals for the largest eigenvalues are distinguishable from the rest. Instead, the mode of convergence for the pairwise distances is weakened to convergence in law. Under the same condition as in Theorem 1, we can derive that as $d \to \infty$,

$$\frac{\|X_1 - X_2\|^2}{d} = \frac{1}{d}(X_1^{\mathrm{T}} X_1 + X_2^{\mathrm{T}} X_2 - 2X_1^{\mathrm{T}} X_2) \Rightarrow \frac{1}{n}(Q + 2c_{k+1})$$

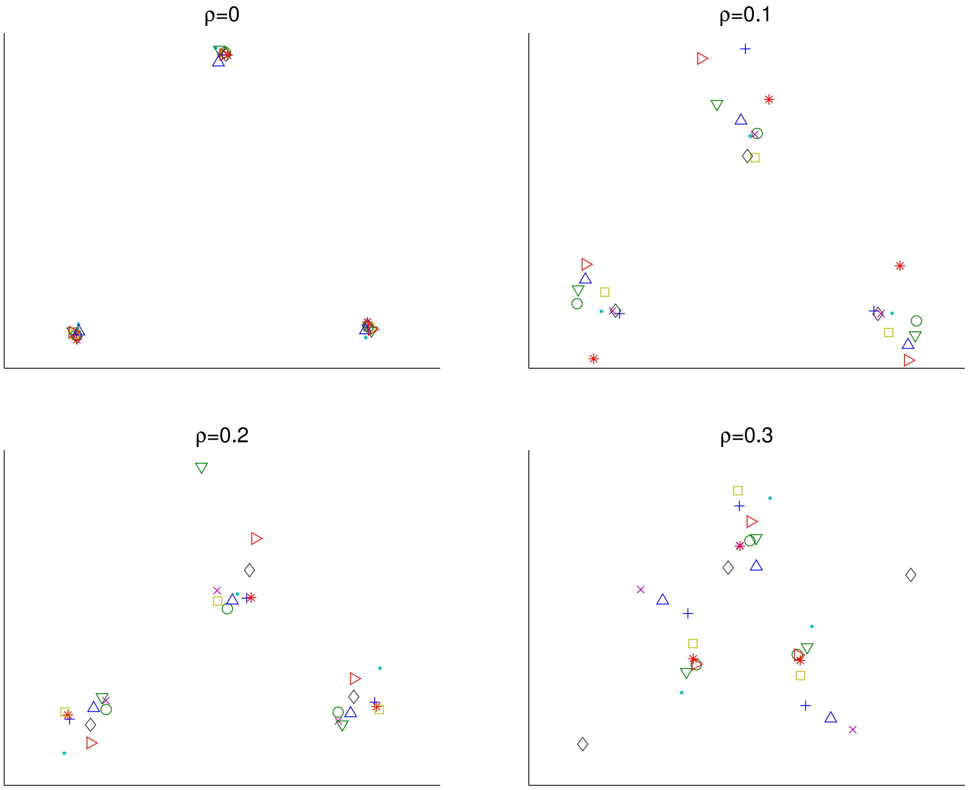

Fig. 1. Geometric representation of 3-sample data from 10 different runs. Data are projected on the hyperplane generated by the data vectors. When $\rho = 0$, the data vectors approximately form an equilateral triangle. As ρ increases, the deterministic structure turns deformalised.

where $Q = w_{11} + w_{22} - 2w_{12}$ with w_{ij} denoting the (i, j)-th entry of $Z_1^T C_1 Z_1$. In contrast to the deterministic limit under the mild spiked model, the pairwise distances between the data vectors under boundary spiked model have a limit distribution.

For an example of the boundary spiked model, consider a compound symmetry (CS) model where all the variables have equal variances, say 1, and the covariance between any of the two variables is $\rho > 0$. The corresponding covariance is $\Sigma = (1 - \rho)I + \rho J$, where I is a $d \times d$ identity matrix and J is a $d \times d$ matrix with all the entries equal to 1. Since the first eigenvalue is $(d - 1)\rho + 1$ and the rest are $(1 - \rho)$, this is a boundary spiked model with $k = 1$, $c_1 = \rho$, and $c_2 = 1 - \rho$. Under the additional Gaussian assumption, we obtain the limiting distribution of the squared distance between the pair of the data vectors as $d \to \infty$,

$$\frac{\|X_1 - X_2\|^2}{d} \Rightarrow \frac{1}{n}\{2\rho\chi_1^2 + 2(1 - \rho)\},$$

where χ_n^2 represents a chi-square distribution with the degrees of freedom n. We generate Gaussian random samples of sample size $n = 3$ and $d = 1000$ with varying covariances $\rho = 0, 0.1, \ldots, 0.3$. In each panel of Fig. 1, the triangle formed by the 3-data vectors is projected so that the line formed by the two points on the bottom is parallel to the x-axis and the middle of the two is located on the origin. To show the sampling variation, triangles from 10 replications are presented. The $\rho = 0$ case corresponds to a spherical Gaussian and the 3-data vectors in the large dimensional space form an approximately equilateral triangle. As the correlation grows, the 3-data points are more deviated from the equilateral triangle.

4. The HDLSS asymptotic behavior of PCA

In this section, we study the large d asymptotic behavior of the sample eigenvalues and the eigenvectors at the boundary spiked model. In the extreme spiked population model when the first k eigenvalues grow at the rate of $O(d^\alpha)$ with $\alpha > 1$, i.e., the assumption (A2) is replaced by

(A2′) $\lambda_{i,d}/d^\alpha \to c_i$ as $d \to \infty$ for some $\alpha > 1$ and $c_i > 0$, $i = 1, \ldots, k$.

The limiting distributions of the sample eigenvalues are established by Jung and Marron [13]. As $d \to \infty$,

$$\begin{cases} \dfrac{\hat{\lambda}_i}{d^\alpha} \Rightarrow \dfrac{1}{n}\varphi_i(Z_1^{\mathrm{T}} C_1 Z_1) & i = 1, \ldots, k \\ \dfrac{\hat{\lambda}_i}{d} \xrightarrow{p} \dfrac{c_{k+1}}{n}, & i = k+1, \ldots, n. \end{cases}$$

The largest k sample eigenvalues converge in distribution at the rate of d^α. They also showed that the corresponding sample eigenvectors are subspace consistent in the sense that the i-th sample PC direction $\hat{v}_i$ lives in the subspace generated by the first k population PC direction vectors $\{v_1, \ldots, v_k\}$,

$$\text{Angle}\ (\hat{v}_i,\ \text{span}\ \{v_1, \ldots, v_k\}) \to 0 \quad \text{as } d \to \infty,\ i = 1, \ldots, k.$$

The rest of the PC directions are strongly inconsistent in the sense that the limiting angle between the j-th sample and population PC directions is $\pi/2$ for $j = k+1, \ldots, n$.

In the boundary spiked model, the limiting distribution of the k largest sample eigenvalues is still at the rate of d, but shifted due to the aggregation effect of the small eigenvalues.

Theorem 2. *Suppose all the assumptions in Theorem 1 hold. Let $\hat{\lambda}_i$ be the sample eigenvalues of the sample covariance matrix S_d and $\varphi_i(A)$ denote the i-th largest eigenvalue of the matrix A. Then, as $d \to \infty$,*

$$\begin{cases} \dfrac{\hat{\lambda}_i}{d} \Rightarrow \dfrac{1}{n}\varphi_i(Z_1^{\mathrm{T}} C_1 Z_1) + \dfrac{c_{k+1}}{n}, & i = 1, \ldots, k \\ \dfrac{\hat{\lambda}_i}{d} \xrightarrow{p} \dfrac{c_{k+1}}{n}, & i = k+1, \ldots, n. \end{cases}$$

Since the largest eigenvalues are mixed with the aggregation of the smallest eigenvalues, it will impact on the consistency of the sample PC directions. In the following proposition, we show that not all of the largest k sample PC directions are subspace consistent on the boundary spiked model.

Proposition 1. *Suppose all the assumptions in Theorem 1 hold. Then, at least one angle between the i-th sample PC direction vector and the subspace spanned by the first k population PC directions is bounded away from 0 in the limit with probability 1, i.e.,*

$$\lim_{d\to\infty} \max_{i=1,\ldots,k} \text{Angle}\ (\hat{v}_i,\ \mathit{span}\ \{v_1, \ldots, v_k\}) > 0.$$

5. Discussion

The actual motivation for this work came from the observation that the structure of the high dimensional data can be important in the binary classification problems. The CS covariance model introduced in Section 3 does not fall in the model category where the HDLSS data representation holds because the data vectors essentially lie in a very low dimensional space, i.e., the subspace generated by the first PC direction vector. This observation becomes crucial when it comes to choose useful classification methods. As noted by Bickel and Levina [7], when the population covariance is CS, a simple independence rule such as Naïve Bayes (NB) does not work well as it works reasonably well for the other types of high dimensional data. Instead, as reported in [1], the maximal data piling (MDP) method outperforms some other sophisticated discrimination methods in this scenario. Data projections onto the MDP direction have no within-class variation. The reason why the seemingly artificial direction of no variation performs well is because it provides a good approximation to the inverse of the direction of the largest variation, i.e., the first PC direction. This in turn means that the MDP method becomes a good generalization of the Fisher's Linear Discrimination method for the HDLSS data. Yet, existing results in the HDLSS asymptotics do not provide good answers as to (1) how far the data representation under CS model is deviated from the regular HDLSS data representation and (2) how strong the low dimensional representation of CS covariance model is observed with the growing dimensionality. Our work on the HDLSS PCA as well as the geometric representation under the boundary model provides good insights to the questions addressed above.

Appendix

Proof of Theorem 1. Using the block matrix notation in (2), the dual sample covariance can be written as

$$nS_D = Z_1^{\mathrm{T}} \Lambda_1 Z_1 + Z_2^{\mathrm{T}} \Lambda_2 Z_2.$$

Then, the weak convergence of the first term on the right-hand is immediate from the assumption (A2),

$$d^{-1} Z_1^{\mathrm{T}} \Lambda_1 Z_1 \Longrightarrow \sum_{i=1}^{k} c_i z_i^{\mathrm{T}} z_i := Z_1^{\mathrm{T}} C_1 Z_1,$$

where $C_1 = \mathrm{diag}\,(c_1, \dots, c_k)$. The probability convergence of the second term follows by applying Theorem 1 in [13] since the assumptions (A1) and (A4) hold,

$$\left(\sum_{i=k+1}^{d} \lambda_{i,d}\right)^{-1} Z_2^{\mathrm{T}} \Lambda_2 Z_2 \to I_n \quad \text{as } d \to \infty.$$

Combining this with (A3), by Slutsky's lemma, we obtain the weak convergence of the dual sample covariance matrix,

$$d^{-1} n S_D \Longrightarrow Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n \quad \text{as } d \to \infty. \quad \square$$

Proof of Theorem 2. Since the eigenvalues are continuous functions of the entries of the matrix, the limit distribution of the eigenvalues follows from Theorem 1,

$$\varphi_i(d^{-1} n S_D) \Rightarrow \varphi_i(Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n), \quad \text{for } i = 1, \dots, n.$$

Let $(s_i, u_i), i = 1, \dots, k$, be the pairs of the eigenvalues and the eigenvectors of $Z_1^{\mathrm{T}} C_1 Z_1$, whose rank is at most k. Since $(Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n) u_i = (s_i + c_{k+1}) u_i$, we conclude that $\{s_i + c_{k+1}, i = 1, \dots, k\}$ are the eigenvalues of the limit matrix, $Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n$. Also, for any vector u that is orthogonal to $\{u_1, \dots, u_k\}$, note that

$$(Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n) u = c_{k+1} u. \tag{A.1}$$

The set of vectors u satisfying (A.1) forms an $(n-k)$-dimensional orthogonal complement of the subspace generated by the row vectors of Z_1, thus, c_{k+1} is the eigenvalue of $Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n$ with the multiplicity $n-k$. Therefore, as $d \to \infty$,

$$\varphi_i(d^{-1} n S_D) \Longrightarrow \varphi_i(Z_1^{\mathrm{T}} C_1 Z_1 + c_{k+1} I_n) = \begin{cases} \varphi_i(Z_1^{\mathrm{T}} C_1 Z_1) + c_{k+1}, & i = 1, \dots, k \\ c_{k+1}, & i = k+1, \dots, n. \end{cases}$$

Combining the results above with the fact that S_d and S_D share the same non-zero eigenvalues completes the proof. $\square$

Proof of Proposition 1. Introduce the standardized sample covariance matrix $\tilde{S}$,

$$\begin{aligned} \tilde{S}_{d \times d} &= \Lambda^{-\frac{1}{2}} V^{\mathrm{T}} S V \Lambda^{-\frac{1}{2}} \\ &= \Lambda^{-\frac{1}{2}} V^{\mathrm{T}} \hat{V} \hat{\Lambda} \hat{V}^{\mathrm{T}} V \Lambda^{-\frac{1}{2}} \\ &= \Lambda^{-\frac{1}{2}} P \hat{\Lambda} P^{\mathrm{T}} \Lambda^{-\frac{1}{2}}, \end{aligned}$$

where $P = \{p_{ij}\} = V^{\mathrm{T}} \hat{V}$. The (i, j)-th entry of P is the inner product between the i-th population PC direction vector and the j-th sample PC direction vector. The inner product between the two direction vectors provides all the information for the angle between them. Thus, we will investigate the limit behavior of the inner product matrix P. For the subspace consistency of the i-th sample PC direction to the subspace spanned by the first k population PC direction vectors, it makes sense to the study the total proportion of the first k population PC direction vectors captured by the i-th sample PC direction vector, $\sum_{j=1}^{k} p_{ji}^2,\ i = 1, \dots, k$. Using the sphered data, the standardized sample covariance matrix can be also expressed as $\tilde{S} = \frac{1}{n} Z Z^{\mathrm{T}}$. From the two representations of $\tilde{S}$, we have the following equality for the (j, j)-th diagonal entry of $\tilde{S}$,

$$\tilde{S}_{jj} = \frac{1}{\lambda_j} \sum_{i=1}^{n} \hat{\lambda}_i p_{ji}^2 = \frac{1}{n} z_j z_j^{\mathrm{T}}, \quad j =, 1 \dots, d. \tag{A.2}$$

By adding the diagonal entries of $\tilde{S}_{jj}$ for $j = 1, \dots, k$, we obtain

$$\begin{aligned} \sum_{j=1}^{k} \frac{1}{n} \lambda_j z_j z_j^{\mathrm{T}} &= \sum_{j=1}^{k} \sum_{i=1}^{n} \hat{\lambda}_i p_{ji}^2 \\ &\geq \sum_{i=1}^{k} \hat{\lambda}_i \cdot \min_{i=1,\dots,k} \left(\sum_{j=1}^{k} p_{ji}^2 \right). \end{aligned}$$

Applying Theorem 2, we obtain the limit of the upper bound for $\min_{i=1,\dots,k} \sum_{j=1}^{k} p_{ji}^2$,

$$\lim_{d\to\infty} \min_{i=1,\dots,k} (p_{1i}^2 + p_{2i}^2 + \cdots + p_{ki}^2) \leq \lim_{d \to \infty} \frac{\operatorname{trace}\,(Z_1^{\mathrm{T}} \Lambda_1 Z_1)/d}{n \sum_{i=1}^{k} \hat{\lambda}_i / d}$$

$$= \frac{\text{trace}\,(Z_1^{\mathrm{T}} C_1 Z_1)}{\sum_{i=1}^{k}\{\varphi_i(Z_1^{\mathrm{T}} C_1 Z_1) + c_{k+1}\}}$$

$$= \frac{\text{trace}\,(Z_1^{\mathrm{T}} C_1 Z_1)}{\text{trace}\,(Z_1^{\mathrm{T}} C_1 Z_1) + k c_{k+1}}.$$

The last term is less than 1 with probability 1, therefore,

$$\lim_{d\to\infty} \max_{i=1,\ldots,k} \text{Angle}\,(\hat{u}_i,\ \text{span}\,\{u_1,\ldots,u_k\}) = \lim_{d\to\infty} \min_{i=1,\ldots,k} \arccos\left(\left(\sum_{j=1}^{k} p_{ji}^2\right)^{1/2}\right) > 0$$

with probability 1. □

References

[1] J. Ahn, J.S. Marron, The maximal data piling direction for discrimination, Biometrika 97 (1) (2010) 254–259.
[2] J. Ahn, J.S. Marron, K.E. Muller, Y.-Y. Chi, The high dimension, low sample size geometric representation holds under mild conditions, Biometrika 3 (94) (2007) 760–766.
[3] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, third ed., Wiley-Interscience, 2003.
[4] Z. Bai, J.W. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, second ed., Springer, 2009.
[5] Z.D. Bai, Y.Q. Yin, Limit of the smallest eigenvalue of a large dimensional sample covariance matrix, The Annals of Probability 21 (3) (1993) 1275–1294.
[6] J. Baik, J.W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, Journal of Multivariate Analysis 97 (6) (2006) 1382–1408.
[7] P. Bickel, E. Levina, Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations, Bernoulli 10 (6) (2004) 989–1010.
[8] J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using a factor model, Journal of Econometrics 147 (1) (2008) 186–197.
[9] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Statistica Sinica 20 (1) (2010) 101–148.
[10] P. Hall, J.S. Marron, A. Neeman, Geometric representation of high dimension low sample size data, The Journal of the Royal Statistical Society, Series B 67 (3) (2005) 427–444.
[11] I.M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, The Annals of Statistics 29 (2) (2001) 295–327.
[12] I.T. Jolliffe, Principal Component Analysis, second ed., Springer, 2002.
[13] S. Jung, J.S. Marron, PCA consistency in high dimension, low sample size context, The Annals of Statistics 37 (6B) (2009) 4104–4130.
[14] Y. Liu, D.N. Hayes, A. Nobel, J.S. Marron, Statistical significance of clustering for high-dimension, low-sample size data, Journal of the American Statistical Association 103 (483) (2008) 1281–1293.
[15] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, Statistica Sinica 17 (4) (2007) 1617–1642.
[16] S. Portnoy, Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency, The Annals of Statistics 12 (4) (1984) 1298–1309.
[17] X. Qiao, H.H. Zhang, Y. Liu, M.J. Todd, J.S. Marron, Weighted distance weighted discrimination and its asymptotic properties, Journal of the American Statistical Association 105 (489) (2010) 401–414.